AMENDMENTS TO THE CLAIMS

The listing of claims will replace all prior versions, and listings of claims in the application:

LISTING OF CLAIMS:

1. (Currently amended) A method for computing a measure of similarity between a first (or input) document and one or more disparate (or search results) documents, comprising:

(a) receiving a first document and identifying the best keywords in the text by recognizing rare and uncommon keywords, including keywords that belong to one or more domain specific or subject matter specific dictionary;

(b) identifying documents similar to the first document using a query by formulating wrappers using the list of the best keywords identified in the first document that also appear in a DS dictionary;

(c) receiving a first list of rated keywords extracted from the first document and a list of rated keywords extracted from each of the one or more disparate documents;

(d) comparing the first list of rated keywords to the list of rated keywords from each of the one or more disparate documents to determine whether the first document forms part of the one or more disparate documents using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the list of at least one of the one or more disparate documents;

(e) verifying inclusion of the first document in the one or more disparate documents by computing a second percentage for each of the one or more disparate documents indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents when the first computed percentage indicates that the first document is included in at least one of the one or more disparate documents;

(f) using the first computed percentage to specify the measure of similarity when the computed second percentage for at least one of the one or more disparate documents is greater than the first computed percentage;

(g) ranking the one or more disparate documents based on the percentage computed indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents;

(h) if the first computed percentage does not indicate that the first document is included in the second document, computing a third percentage using the Jaccard similarity distance measure, wherein if said Jaccard similarity distance measure is greater than about 90 percent, the second document is identified as a revision of the first document, and if the Jaccard similarity distance measure is less than about 90 percent, said measure is a similarity measure between said first and second document; and

(i) if the third computed percentage indicates that the first document is a revision of the second document, computing a fourth percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the second list also exist in the first list.

2. (Original) The method according to claim 1, wherein the second percentage at (c) is computed by giving weight only to those keywords and their set of neighboring keywords in the first list that match in the second list and a threshold percentage of the keywords in their set of neighboring keywords.

3. (Original) The method according to claim 2, wherein the second percentage at (c) is computed by giving full weight to those keywords in the first list of rated keywords that cannot be accurately identified as having a complete set of neighboring keywords in the second set of keywords.

4. (Original) The method according to claim 2, wherein the threshold percentage is reduced when the first list of rated keywords is identified using OCR.

5. (Cancelled).

6. (Cancelled).

7. (Currently Amended). The method according to claim ~~6~~ 1, further comprising the fourth computed percentage to specify the measure of similarity except when: (i) the fourth computed percentage is greater than the second computed percentage; (ii) the first list of rated keywords is identified using OCR; (iii) the fourth computed percentage is greater than fifty percent; and (iv) less than twenty percent of the keywords in the first list of keywords are in the second list of keywords.

8. (Original) The method according to claim 1, wherein the first computed percentage indicates that the first document is included in the second document when the percentage defined by ratio of Sum1/Sum2 is greater than approximately ninety percent, where: D1 is the number of keywords in first list of keywords; D2 is the number of keywords in the second list of keywords; Sum1 is the sum of the weights of keywords that appear in D1 that also appear in D2; Sum2 is the sum of the weights of keywords in D1.

9. (Original) The method according to claim 1, wherein the first list of rated keywords includes one or more keywords translated from a second language different from a first language that is identified as being a primary language of the first document.

10. (Original) The method according to claim 1, wherein the first document is a portion of the second document.

11. (Currently amended) A computer-based system for computing a measure of similarity between a first (or input) document and one or more (or search results) documents, comprising:

(a) means for receiving a first document and identifying the best keywords in the text by recognizing rare and uncommon keywords, including keywords that belong to one or more domain specific or subject matter specific dictionary;

(b) means for identifying documents similar to the first document using a query by formulating wrappers using the list of the best keywords identified in the first document that also appear in a DS dictionary;

(c) means for receiving a first list of rated keywords extracted from the first document and a list of rated keywords extracted from each of the one or more disparate documents, wherein keywords are rated at least in part by a relevant weight from their associated document language;

(d) means for comparing the first list of rated keywords to the list of keywords from each of the one or more disparate documents to determine whether the first document forms part of the one or more disparate documents using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the list of at least one of the one or more disparate documents;

(e) means for verifying inclusion of the first document in the one or more disparate documents by computing a second percentage for each of the one or more disparate documents indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents when the first computed percentage indicates that the first document is included in at least one of the one or more disparate documents; and

(f) means for using the first computed percentage to specify the measure of similarity when the computed percentage for at least one of the one or more disparate documents is greater than the first computed percentage;

(g) means for ranking the one or more disparate documents based on the percentage computed indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the list for at least one of the one or more disparate documents;

(h) if the first computed percentage does not indicate that the first document is included in the second document, means computes a third percentage using the Jaccard distance measure, wherein if said Jaccard similarity distance measure is greater than about 90 percent, the second document is identified as a revision of the first document, and if the Jaccard similarity distance measure is less than about 90

percent, said measure is a similarity measure between said first and second document; and

(i) if the third computed percentage indicates that the first document is a revision of the second document, means computes a fourth percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the second list also exist in the first list.

12. (Original) The system according to claim 11, wherein the second percentage at (c) is computed by said computing means by giving weight only to those keywords and their set of neighboring keywords in the first list that match in the second list and a threshold percentage of the keywords in their set of neighboring keywords.

13. (Original) The system according to claim 12, wherein the second percentage at (c) is computed by said computing means by giving full weight to those keywords in the first list of rated keywords that cannot be accurately identified as having a complete set of neighboring keywords in the second set of keywords.

14. (Original) The system according to claim 12, wherein the threshold percentage is reduced when the first list of rated keywords is identified using OCR.

15. (Cancelled)

16. (Cancelled)

17. (Original) The system according to claim 16, further comprising means for using the fourth computed percentage to specify the measure of similarity except when: (i) the fourth computed percentage is greater than the second computed percentage; (ii) the first list of rated keywords is identified using OCR; (iii) the fourth computed percentage is greater than fifty percent; and (iv) less than twenty percent of the keywords in the first list of keywords are in the second list of keywords.

18. (Original) The system according to claim 11, wherein the first computed percentage indicates that the first document is included in the second document when the percentage defined by ratio of Sum1/Sum2 is greater than approximately ninety percent, where: D1 is the number of keywords in first list of keywords; D2 is the number of keywords in the second list of keywords; Sum1 is the sum of the weights of keywords that appear in D1 that also appear in D2; Sum2 is the sum of the weights of keywords in D1.

19. (Original) The system according to claim 11, wherein the first list of rated keywords includes one or more keywords translated from a second language different from a first language that is identified as being a primary language of the first document.

20. (Currently Amended) An article of manufacture for computing a measure of similarity between a first (or input) document and one or more disparate (or search results) documents, the article of manufacture comprising computer usable media including computer readable instructions embedded therein that causes a computer to perform a method, wherein the method comprises:

(a) receiving a first document and identifying the best keywords in the text by recognizing rare and uncommon keywords, including keywords that belong to one or more domain specific or subject matter specific dictionary;

(b) identifying documents similar to the first document using a query by formulating wrappers using the list of the best keywords identified in the first document that also appear in a DS dictionary;

(c) receiving a first list of rated keywords extracted from the first document and a second list of rated keywords extracted from the second document;

(d) using the first and second lists of rated keywords to determine whether the first document forms part of the second document using a first computed percentage indicating what percentage of keyword ratings in the first list also exist in the second list;

(e) verifying inclusion of the first document in the second document computing a second percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the first list also exist in the second list when the

first computed percentage indicates that the first document is included in the second document;

(f) using the first computed percentage to specify the measure of similarity when the second computed percentage is greater than the first computed percentage;

(g) if the first computed percentage does not indicate that the first document is included in the second document, computing a third percentage using the Jaccard distance measure, wherein if said Jaccard similarity distance measure is greater than about 90 percent, the second document is identified as a revision of the first document, and if the Jaccard similarity distance measure is less than about 90 percent, said measure is a similarity measure between said first and second document; and

(h) if the third computed percentage indicates that the first document is a revision of the second document, computing a fourth percentage indicating what percentage of keyword ratings along with a set of their neighboring keyword ratings in the second list also exist in the first list, the fourth computed percentage is used to specify the measure of similarity except when: (i) the fourth computed percentage is greater than the second computed percentage; (ii) the first list of rated keywords is identified using OCR; (iii) the fourth computed percentage is greater than fifty percent; and (iv) less than twenty percent of the keywords in the first list of keywords are in the second list of keywords.